Gene Expression Data Analysis Using Heuristic Attribute Reduction in Rough Set Theory

doi: info:doi/10.14864/softscis.2010.0.1574.0

Gene Expression Data Analysis Using Heuristic Attribute Reduction in Rough Set Theory

doi: info:doi/10.14864/softscis.2010.0.1574.0

Gene Expression Data Analysis Using Heuristic Attribute Reduction in Rough Set Theory

Yasuo Kudo
College of Information and Systems
Muroran Institute of Technology
27-1 Mizumoto, Muroran 050-8585, Japan
Email: kudo@csse.muroran-it.ac.jp

Yoshifumi Okada
College of Information and Systems
Muroran Institute of Technology
27-1 Mizumoto, Muroran 050-8585, Japan
Email: okada@csse.muroran-it.ac.jp

***Abstract*—We apply a combined method of heuristic attribute reduction and evaluation of relative reducts in rough set theory to gene expression data analysis. Our method extracts as many relative reducts as possible from gene expression data and select the best relative reduct from a viewpoint of constructing useful decision rules. As an experimental result, our method extracted decision rules about a gene that has been identified as a novel biomarker of human breast cancer in recent studies. This result indicates a possibility of our method as a useful tool for gene expression data analysis.**

I. Introduction

DNA microarray technology has enabled us to monitor expression levels of thousands of genes simultaneously under certain condition and has been yielded various applications in the field of disease diagnosis [23], drug discovery [4] and toxicological research [20]. Among them, cancer informatics based on gene expression data is an important domain which has promising prospects for both clinical treatment and biomedical research. One of the key issues in this domain is to discover biomarker genes for cancer diagnosis from massive gene expression data using bioinformatics approaches called a gene selection.

Typical gene selection approach is a statistical test such as t-test and ANOVA [2]. This approach detects differentially expressed genes between samples from different cells/tissues. Most of the statistical tests assume that expression values of each gene over the samples follow a prior probability distribution; hence sufficient large number of samples is required for obtaining statistically reliable results.

Rough set theory [15], [16] provides a theoretical basis of set-theoretical approximation and rule generation from categorical data. Attribute reduction is one of the most important and hot research topics in rough set theory as a basis of rule generation by rough set theory and there have been many proposals of heuristic algorithms to compute some candidates of relative reducts (for example, see [3], [5], [7], [10], [11], [22], [24]). Kudo and Murai proposed attribute reduction algorithms to compute as many relative reducts as possible from a decision table with numerous condition attributes [9]. They also proposed an evaluation criterion of relative reducts that evaluates usefulness of relative reducts from a viewpoint of decision rule generation [8].

In this paper, we introduce Kudo and Murai's heuristic attribute reduction algorithms [9] and a criterion of relative reducts [8] to gene expression data analysis. We use these algorithms and criterion to two gene expression datasets: breast cancer [21] and leukemia [1] and discuss about the extracted decision rules from these datasets and its biological meanings. Experimental results indicate that the method used in this paper can identify differentially expressed genes between different classes in gene expression datasets and useful for gene expression data analysis.

The rest of this paper is organized as follows. In Section II, we review about gene expression datasets obtained by DNA microarray experiments. In Section III, we introduce a heuristic method of attribute reduction [9] and an evaluation criterion of relative reducts [8] as methods we used in this paper for generating decision rules from gene expression datasets. We apply the methods to two gene expression datasets and discuss the experimental results in Section IV, and concludes this paper in Section V. In Appendix, we review Pawlak's rough set theory as the background of this paper.

II. Gene expression data

Figure 1 illustrates a gene expression dataset obtained by DNA microarray experiments. A single DNA microarray can measure expression levels of thousands of genes simultaneously in cells/tissue under a certain condition (called a sample). Each point on a DNA microarray indicates one kind of a gene, and its intensity represents the expression level. These intensities are converted into numerical data. In multiple DNA microarray experiments, a gene expression dataset is provided by the form of a matrix as shown in Fig. 1, in which each row and each column correspond to a sample and a gene, respectively, and each element is an expression value of a gene.

III. Methods

Methods that we use in this paper to extract decision rules from gene expression data based on rough set theory consist of the following three parts:

1) Extraction of relative reducts from gene expression data.
2) Selection of relative reducts in accordance with an evaluation criterion of relative reducts.

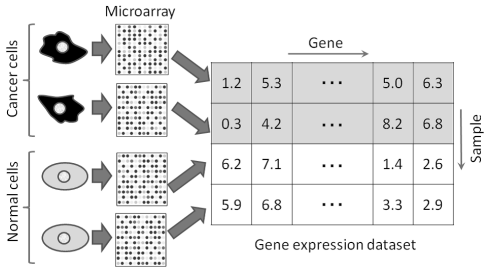

Fig. 1. Gene expression dataset obtained by DNA microarray experiments

3) Construction of decision rules from the selected relative reducts.

Below, as the methods we use in this paper, we introduce a heuristic attribute reduction algorithm for generating as many relative reducts as possible [9] and a criterion for evaluating usefulness of relative reducts [8]. Note that the details of rough set theory as the background of these algorithms and criterion are in Appendix.

A. Heuristic Algorithm for Attribute Reduction Using Reduced Decision Tables

In this section, we review a heuristic algorithm to generate as many relative reducts as possible from decision tables with numerous condition attributes proposed by Kudo and Murai [9].

This heuristic algorithm is based on the idea of reduced decision tables that preserve the discernibility of objects that belong to mutually different decision classes in the given decision table. Formally, a reduced decision table of a given decision table is defined as follows.

Definition 1: Let $DT = (U, C, d)$ be a decision table. A reduced decision table of DT is the following triple:

$$RDT = (U, C', d), \tag{1}$$

where U and d are identical to DT. The set of condition attributes C' satisfies the following conditions:

1) $C' \subseteq C$.
2) For any objects x_i and x_j that belong to different decision classes, if x_i and x_j are discernible by R_C, x_i and x_j are also discernible by $R_{C'}$.

Algorithm 1 below generates a reduced decision table of the given decision table. In Algorithm 1, condition attributes are selected from C at random based on the parameter of base size b that decides the minimum number of condition attributes of the reduced decision table, and supply some attributes in elements of the discernibility matrix to preserve discernibility of objects in the given decision table.

Note that, for any decision table and any reduced decision table, a set of condition attributes A is a relative reduct of the reduced decision table if and only if A is also a relative reduct of the given decision table [9]. Thus generating as many reduced decision tables and those relative reducts as possible from the given decision table, we can extract many relative reducts from the given decision table. The following algorithm

Algorithm 1 dtr: decision table reduction algorithm

Input: decision table $DT = (U, C, d)$,
discernibility matrix DM of DT,
base size b
Output: reduced decision table (U, C', d)

```
Select b attributes a_1, ..., a_b from C at random by sampling without replacement
C' = {a_1, ..., a_b}
for all δ_ij ∈ DM such that i > j do
    if δ_ij ≠ ∅ and δ_ij ∩ C' = ∅ then
        Select c ∈ δ_ij at random
        C' = C' ∪ {c}
    end if
end for
return (U, C', d)
```

Algorithm 2 below generates relative reducts of the given decision table based on generating reduced decision tables by Algorithm 1 and switching exhaustive attribute reduction and heuristic attribute reduction according to the number of condition attributes of each reduced decision table.

Algorithm 2 Exhaustive / heuristic attribute reduction

Input: decision table $DT = (U, C, d)$,
base size b, size limit L, number of iteration I
Output: set of candidates of relative reduct RED

```
RED = ∅
DM ← the discernibility matrix of DT
if |C| ≤ L then
    RED ← result of exhaustive attribute reduction from DT
else
    for i = 1 to I do
        RDT = dtr(DT, DM, b)
        if |C'| ≤ L then
            S ← result of exhaustive attribute reduction from RDT
        else
            S ← result of heuristic attribute reduction from RDT
        end if
        RED = RED ∪ S
    end for
end if
return RED
```

In Algorithm 2, the size limit L is the threshold for switching attribute reduction methods and if the number of condition attributes of a decision table is smaller than L, Algorithm 2 tries to generate the set of all relative reducts of the decision table. Thus, we need to set the threshold L appropriately. If the number of condition attributes of the given decision table DT is grater than the threshold L, Algorithm 2 repeats I times of generating a reduced decision table RDT

and attribute reduction from RDT by selecting the exhaustive method or the heuristic method, and generate the set RED of relative reducts. Note that RED may contain some output with redundancy if the result of the heuristic attribute reduction is not guaranteed to generate relative reducts.

B. An evaluation criterion of relative reducts

From the viewpoint of data analysis using rough set theory, Kudo and Murai [8] proposed an evaluation criterion of relative reducts for extracting useful decision rules. This criterion evaluates usefulness of each relative reduct by average of coverage of decision rules generated from the relative reduct. Let $DT = (U, C, d)$ be a decision table. For any non-empty set $B \subseteq C$ of condition attributes, the average of coverage $ACov(B)$ of all decision rules generated from B is calculated as follows [8]:

$$ACov(B) = \frac{|\mathcal{D}|}{\sum_{[x]_B \in U/R_B} |\{D_j \in \mathcal{D} \mid [x]_B \cap D_j \neq \emptyset\}|} \quad (2)$$

$$= \frac{\text{Num. of decision classes}}{\text{Num. of rules generated from } B}, \quad (3)$$

where U/R_B is the quotient set of U by the equivalence relation R_B and $|X|$ is the cardinality of the set X. This equation indicates that $ACov(B)$ depends only on the number of decision rules because the number of decision classes is fixed in the given decision table.

For each relative reduct $E \subseteq C$ of the given decision table, the average of coverage $ACov(E)$ reflects roughness of partition U/R_E by equivalent classes based on R_E. It is guaranteed that, for any relative reducts $E, F \subseteq C$, if the partition U/R_E is rougher than the partition U/R_F, then $ACov(E) \geq ACov(F)$ holds and this property provides a theoretical basis for using Eq. (2) as an evaluation criterion of relative reducts [8].

IV. Experiments and discussion

A. Datasets and preprocessing

To evaluate the usefulness of our method, we use two gene expression datasets: breast cancer [21] and leukemia [1]. Both of them are two-class dataset. The leukemia dataset is composed of gene expression values for 12,582 genes in 24 Acute Lymphocytic Leukemia (ALL) samples and 28 Acute Myeloid Leukemia (AML) samples. The breast cancer dataset includes gene expression values for 7,129 genes in 25 positive and 24 negative samples. For each dataset, the expression values from each gene are linearly normalized to have mean 0 and variance 1. Subsequently, they are discretized to six bins $(-3, -2, -1, 1, 2, 3)$ by uniformly dividing the difference between the maximum and the minimum in the normalized data and one bin that represents lack of gene expression values. Discretized positive values represent that the genes with positive values are up-regulated, while negative values represent that genes are down-regulated.

B. Procedures and results of the experiments

In the experiments, firstly we used the heuristic attribute reduction algorithms [9] to each dataset with the following parameters; the base size $b = 10$, the size limitation $L = 25$, and the number of iterations $I = 100$ and generated relative reducts of each dataset. Next, for each dataset, we selected the best relative reduct of each dataset in the sense of the evaluation criterion of relative reducts [8]. Finally, we extracted decision rules from each dataset by the following three steps: 1) generating all decision rules by the best relative reduct of each dataset, 2) removing decision rules that contain null values in the antecedents, and 3) combining the generated decision rules as long as possible by interpreting the meanings of decision rules.

As the results, we got the following decision rules for each dataset.

The breast cancer dataset:

1) (*CRIP1* ≥ -2) $\rightarrow$ (class = Positive), Certainty = 0.76, Coverage = 0.64.
2) (*CRIP1* $= -3$) $\rightarrow$ (class = Negative), Certainty = 0.95, Coverage = 0.79.

The leukemia dataset:

3) (*POU2AF1* ≥ -2) $\rightarrow$ (class = ALL), Certainty = 1.0, Coverage = 0.88.
4) (*POU2AF1* $= -3$) $\rightarrow$ (class = AML), Certainty = 1.0, Coverage = 1.0.

C. Extracted rules and their biological meanings

Below, these extracted rules are evaluated on the basis of known biological findings. To this end, we investigate the functions of genes in the rules by reference to a genetic disease database (OMIM) [14] and a protein sequence database (Swiss-Prot) [18].

For breast cancer dataset, the samples can be discriminated into a true class with an accuracy of 95 percent according to the expression level of Cystein-rich intestinal protein 1 (*CRIP1*). *CRIP1* is a transcription factor gene that induces apoptosis in cancer cells. Interestingly, this gene has been identified as a novel biomarker of human breast cancer in recent studies [13], [12]. In the extracted rule, we can see that *CRIP1* expression is more up-regulated in the positive samples. Indeed, this is consistent with the recent findings by Ma et al. [13] that *CRIP1* in human breast cancer was overexpressed compared to normal breast tissue by in situ experiments.

For leukemia dataset, all samples can be perfectly discriminated by the expression level of POU class 2 associating factor 1 (*POU2AF1*). *POU2AF1* is known as a gene responsible for leukocyte differentiation. In Swiss-Prot, we can see the description that "a chromosomal aberration involving *POU2AF1* may be a cause of a form of B-cell leukemia." Namely, it suggests that this gene can be inactivated/down-regulated in lymphocytic leukemia such as ALL. In contrast, it should be noted that *POU2AF1* in the extracted rule shows weaker expression in AML than ALL. At present, detailed role of *POU2AF1* in AML has not been revealed [6] whereas we

expect that biological relevance will be unveiled by experimental biologists in the near future.

V. Conclusion

In this paper, we introduced a combine method of heuristic attribute reduction and evaluation of relative reducts in rough set theory to gene expression data analysis. Our method is based on a heuristic attribute reduction algorithm for generating as many relative reducts as possible [9] and a criterion for evaluating usefulness of relative reducts [8]. We applied our method to two gene expression datasets: breast cancer [21] and leukemia [1]. Experimental results showed that our method can identify differentially expressed genes between different classes in gene expression datasets. For the breast cancer dataset, our method extracted decision rules about a gene that has been identified as a novel biomarker of human breast cancer in recent studies [13], [12]. For the leukemia dataset, decision rules about a gene responsible for leukocyte differentiation were extracted. Thus, these results indicate a possibility of our method as a useful tool for gene expression data analysis.

For future issues, we aim to apply our method to various gene expression datasets and compare our method with other gene expression analysis methods.

Appendix

As the theoretical background of this paper, we review the rough set theory, in particular, decision tables, relative reducts, and discernibility matrices. Note that the contents of this section are based on [9], [17].

A. Decision table and lower and upper approximations

Generally, data analysis subjects by rough sets are described by decision tables. Formally, a decision table is characterized by the following triple:

$$DT = (U, C, d), \tag{4}$$

where U is a finite and nonempty set of objects, C is a finite and nonempty set of condition attributes, and d is a decision attribute such that $d \notin C$. Each attribute $a \in C \cup \{d\}$ is a function $a : U \to V$, where V is a set of values of attributes.

Indiscernibility relations based on subsets of attributes provide classifications of objects in decision tables. For any set of attributes $A \subseteq C \cup \{d\}$, the indiscernibility relation R_A is the following binary relation on U:

$$R_A = \{(x, y) \mid a(x) = a(y), \forall a \in A\}. \tag{5}$$

If a pair (x, y) is in R_A, then two objects x and y are indiscernible with respect to all attributes in A. It is well-known that any indiscernibility relation is an equivalence relation and equivalence classes by an equivalence relation consist of a partition on the domain of the equivalence relation. In particular, the indiscernibility relation R_d based on the decision attribute d provides a partition $\mathcal{D} = \{D_1, \cdots, D_k\}$, and each element $D_i \in \mathcal{D}$ is called a decision class.

Classifying objects with respect to condition attributes provides set-theoretical approximation of decision classes. Formally, for any set $B \subseteq C$ of condition attributes and any decision class $D_i \in \mathcal{D}$, we let:

$$\underline{B}(D_i) \quad = \quad \{x \in U \mid [x]_B \subseteq D_i\}, \tag{6}$$

where the set $[x]_B$ is the equivalence class of x by the indiscernibility relation R_B. The set $\underline{B}(D_i)$ is called the lower approximation of the decision class D_i with respect to B.

B. Relative reducts and discernibility matrices

By checking values of all condition attributes, we can classify all discernible objects of the given decision table to the corresponding decision classes. However, not all condition attributes may need to be checked in the sense that some condition attributes are essential to classify and the other attributes are redundant. A minimal set of condition attributes to classify all discernible objects to correct decision classes is called a relative reduct of the decision table.

Formally, a set $A \subseteq C$ is called a relative reduct of the decision table DT if the set A satisfies the following conditions:

1) $\mathrm{POS}_A(\mathcal{D}) = \mathrm{POS}_C(\mathcal{D})$.
2) $\mathrm{POS}_B(\mathcal{D}) \neq \mathrm{POS}_C(\mathcal{D})$ for any proper subset $B \subset A$,

where $POS_X(\mathcal{D}) \stackrel{\mathrm{def}}{=} \bigcup_{D_i \in \mathcal{D}} \underline{X}(D_i)$ is a set of objects that are correctly classified to those decision classes by checking all attributes in $X \subseteq C$ and called the positive region of $\mathcal{D}$ by X. Note that, in general, there are plural relative reducts in a decision table. The common part of all relative reducts is called the core of the decision table.

The discernibility matrix [19] is one of the most popular methods to compute all relative reducts in the decision table. Let DT be a decision table with $|U|$ objects, where $|U|$ is the cardinality of U. The discernibility matrix DM of DT is a symmetric $|U| \times |U|$ matrix whose element at i-th row and j-th column is the following set of condition attributes to discern between two objects x_i and x_j:

$$\delta_{ij} = \begin{cases} \{a \in C \mid a(x_i) \neq a(x_j)\}, & \text{if } d(x_i) \neq d(x_j) \text{ and} \\ & \{x_i, x_j\} \cap \mathrm{POS}_C(\mathcal{D}) \neq \emptyset, \\ \emptyset, & \text{otherwise.} \end{cases} \tag{7}$$

Each element $a \in \delta_{ij}$ represents that x_i and x_j are discernible by checking the value of a.

C. Certainty and coverage of decision rules

Certainty and coverage are well-known criteria of decision rules for evaluating accuracy and relevance of decision rules, respectively. Formally, the certainty and the coverage of a decision rule constructed from a set $B \subseteq C$ of condition attributes, the decision attribute d, and an object $x \in U$ are defined by

$$\text{Certainty} \quad = \quad \frac{|[x]_B \cap D_i|}{|[x]_B|}, \tag{8}$$

$$\text{Coverage} \quad = \quad \frac{|[x]_B \cap D_i|}{|D_i|}, \tag{9}$$

where $D_i \in \mathcal{D}$ is the decision class such that $x \in D_i$.

REFERENCES

[1] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, S. J. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, Vol. 30, No. 1, pp. 41–47, 2002.

[2] X. Cui, G. A. Churchill, Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*, Vol. 4, No. 4, p. 210, 2003.

[3] A. Chouchoulas and A. Shen, Rough Set-Aided Keyword Reduction for Text Categorization, *Applied Artificial Intelligence*, Vol. 15, No. 9, pp.843–873, 2001.

[4] C. Debouck, P. N. Goodfellow, DNA microarrays in drug discovery and development. *Nature Genetics*, Vol. 21, No. 1s, pp. 48–50, 1999.

[5] J. W. Guan and D. A. Bell, Rough computational methods for information systems, *Artificial Intelligence*, Vol. 105, pp.77–103, 1998.

[6] S. E. Gibson, H. Y. Dong, A. S. Advani, E. D. Hsi, Expression of the B cell-associated transcription factors PAX5, OCT-2, and BOB.1 in acute myeloid leukemia: associations with B-cell antigen expression and myelomonocytic maturation. *American Journal of Clinical Pathology*, Vol. 126, No. 6, pp. 916–924, 2006.

[7] K. Hu, L. Diao, Y. Lu, and C. Shi, A Heuristic Optiaml Reduct Algorithm, *Proc. of IDEAL2000*, LNSC 1983, pp.139–144, Springer, 2000.

[8] Y. Kudo and T. Murai, An Evaluation Method of Relative Reducts Based on Roughness of Partitions, *International Journal of Cognitive Informatics and Natural Intelligence*, Vol. 4, No. 2, pp.50–62, 2010.

[9] Y. Kudo and T. Murai, An Attribute Reduction Algorithm by Switching Exhaustive and Heuristic Computation of Relative Reducts, *Proc. of 2010 IEEE International Conference on Granular Computing*, IEEE, pp.265–270, 2010.

[10] Y. Kudo and T. Murai, Heuristic Algorithm for Attribute Reduction Based on Classification Ability by Condition Attributes, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 15, No. 1, to appear.

[11] M. Kryszkiewicz and P. Lasek, FUN: Fast Discovery of Minimal Sets of Attributes Functionally Determining a Decision Attribute, *Transactions on Rough Sets IX*, LNCS 5390, pp.76–95, 2008.

[12] S. Liu, A. Stromberg, H. Tai, J. A. Moscow, Thiamine transporter gene expression and exogenous thiamine modulate the expression of genes involved in drug and prostaglandin metabolism in breast cancer cells. *Molecular Cancer Research*, Vol. 2, No. 8 pp. 477–487, 2004.

[13] X. J. Ma, R. Salunga, J. T. Tuggle, J. Gaudet, P. McQuary et al. Gene expression profiles of human breast cancer progression. *Proceedings of the National Academy of Sciences*, Vol. 100, No. 10, pp. 5974–5979, 2003.

[14] http://www.nslij-genetics.org/search_omim.html

[15] Z. Pawlak, Rough Sets, *International Journal of Computer and Information Science*, Vol. 11, pp.341–356, 1982.

[16] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publisher, 1991.

[17] L. Polkowski, *Rough Sets: Mathematical Foundations*, Advances in Soft Computing, Physica-Verlag, 2002.

[18] http://au.expasy.org/sprot/

[19] A. Skowron and C. M. Rauszer, The discernibility matrix and functions in information systems, *Intelligent Decision Support: Handbook of Application and Advance of the Rough Set Theory*, Słowiński, R. (ed.), Kluwer Academic Publishers, pp.331–362, 1992.

[20] K. E. Vrana, W. M. Freeman, M. Aschner, Use of microarray technologies in toxicology research. *Neurotoxicology*, Vol. 24, No. 3, pp. 321–32, 2003.

[21] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson Jr, J. R. Marks, J. R. Nevins, Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*, Vol. 98, No. 20, pp. 11462–11467, 2001.

[22] Y. Y. Yao, Y. Zhao, J. Wang, and S. Han, A Model of User-Oriented Reduct Construction for Machine Learning, *Transactions on Rough Sets VIII*, LNCS 5084, pp.332–351, 2008.

[23] S. M. Yoo, J. H. Choi, S. Y. Lee, N. C. Yoo, Applications of DNA microarray in disease diagnostics. *Journal of Microbiology and Biotechnology*. Vol. 19, No. 7, pp. 635–46, 2009.

[24] J. Zhang, J. Wang, D. Li, H. He, and J. Sun, A New Heuristic Reduct Algorithm Based on Rough Sets Theory, *Proc. of WAIM2003*, LNCS 2762, pp.247–253, 2003.